A Survey on Named Entity Recognition

Yan Wen[1(✉)], Cong Fan[1], Geng Chen[2], Xin Chen[2], and Ming Chen[3]

[1] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
wenyan84@hotmail.com
[2] College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[3] State Grid Shandong Electric Power Company, Qingdao Power Supply Company, Qingdao 266500, China

Abstract. Natural language processing is an important research direction and research hotspot in the field of artificial intelligence. Named entity recognition is one of the key tasks, which is to identify entities with specific meanings in the text, such as names of people, places, institutions, proper nouns, etc. Traditional named entity recognition methods are mainly implemented based on rules, dictionaries, and statistical learning. In recent years, with the rapid expansion of Internet text data scale and the rapid development of deep learning technology, a large number of deep neural network-based methods have emerged, which have greatly improved the accuracy of recognition. This paper attempts to summarize the traditional methods and the latest research progress in the field of named entity identification, and summarize and analyse its main models, algorithms and applications. Finally, the future development trend of named entity recognition is discussed.

Keywords: Dictionary · CRF · LSTM · Transfer learning · Attention mechanism

1 Introduction

Natural Language Processing (NLP), also known as Information Extraction (IE), is an important research direction in the field of computer science and artificial intelligence. It mainly studies various theories and methods that can realize effective communication between human and computer in natural language. NLP contains multiple subtasks such as word segmentation, named entity recognition, text summarization, machine translation, sentiment analysis, speech recognition and more.

Named Entity Recognition (NER) is one of the key tasks. Named entities were originally proposed at the 6th MUC (Message Understanding Conferences) in 1995 [1], mainly referring to words or phrases with specific names in the text. It generally consists of three major categories (Entity classes, Time classes, and Number class) and seven subclasses (Person name, Place name, Institution name, Time, Date, Currency, and Percentage) [1, 2]. NER is designed to identify these proper nouns in the text and classify them into appropriate categories. NER is the basis for many advanced tasks in

Q. Liang et al. (Eds.): CSPS 2019, LNEE 571, pp. 1803–1810, 2020
https://doi.org/10.1007/978-981-13-9409-6_218

natural language processing, such as syntax analysis, text summary, information retrieval, automatic question and answer and so on.

2 Rule-Based and Dictionary-Based Methods

Rule-based and dictionary-based methods generally analyze the characteristics of an entity and then manually construct rules for matchup, which are the earliest methods used in NER. The rule templates rely on the establishment of knowledge bases and dictionaries [3].

In 1991, Rau [4] published a paper on "Extracting and Identifying Company Names" at the 7th IEEE Artificial Intelligence Application Conference, which mainly used heuristic algorithms and manual writing rules. In 1997, Zhang and Wang [5] used rule-based methods to identify Chinese university names. The accuracy and recall rates were 97.3% and 96.9%, respectively. In 2000, Farmakiotou et al. [6] proposed a rule-based identification method for named entities in Greek financial texts. In 2002, Wang et al. [7] of the Hong Kong Polytechnic University used a rule-based approach for efficient name recognition.

For text with many features, the method of using rules is simple and effective. However, these rules often rely on specific language, domain, etc., so they have poor applicability and high maintenance costs [3].

3 Statistical Learning Based Method

The statistical-based methods build statistical learning models such as Maximum Entropy Models (MEM), Hidden Markov Model (HMM), Conditional Random Field Model (CRF), etc. on the manually labelled corpus and a number of selected features, and the models are used to extract new named entities.

3.1 HMM

HMM is a statistical model that is very widely used and classical. In HMM, $\{x_1, x_2, \ldots, x_{n+1}\}$ is the hidden state of the Markov transition process, $\{y_1, y_2, \ldots, y_{n+1}\}$ is the output, that is, the observation state. The model of the HMM is expressed as:

$$\lambda = (A, B, \pi) \tag{1}$$

where A is the state transition probability matrix; B is the observation probability matrix [8].

In 1999, Bikel et al. [9, 10] proposed an English named entity recognition method based on HMM-IdentiFinder™, and selected number symbols and special character sets as features. For the MUC-6 test text set, the recognition accuracy of English place names, Institution names and Person names reached 97%, 94% and 95%, respectively, and the recall rates could reach 95%, 94% and 94%, respectively. In 2009, the HMM

constructed by Liu [11] could take into account the contextual feature information in the process of word segmentation and labeling, the context to got the best state sequence.

Due to the high efficiency of the Viterbi algorithm, HMM is relatively simple and fast for training. However, due to its output independence assumption, the model has relatively poor performance.

3.2 CRF

CRF is also a classic supervised learning model for sequence labelling. The linear chain conditional random field was the mainstream model of NER before the prevalence of deep neural network based models.

Let $X = (X_1, X_2, \ldots, X_n)$, $Y = (Y_1, Y_2, \ldots, Y_n)$, which are all sequences of random variables represented by linear chains. Given a sequence X of random variables, the conditional probability distribution $P(Y|X)$ of the random variable Y constitutes the CRFs, Markov property is met [8]. The model expression is as follows:

$$P(Y|X) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{2}$$

$$Z(x) = \sum_{Y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{3}$$

where t_k and s_l are characteristic functions; λ_k and μ_l are corresponding weights; $Z(x)$ is a normalization factor.

In 2006, Krishnan and Manning [12] used the CRF model to extract global constraint information from NER. The paper trained two CRF models: one that used local features for prediction; the other extracted local information and features from the output of the previous CRF model. The F1 value of the second CRF model reached 87.24%.

The performance of the CRF algorithm is better than that of the HMM. It does not normalize every node, but all features are globally normalized, so the global optimal value can be obtained. But the training of CRF model is more complicated, with slow convergence and long training time.

There are other methods of statistical-based that are not mentioned as follows. In 1998, Borthwick et al. [13] proposed MEM-based NER method for English and Japanese. Sekine et al. [14] proposed a Japanese-named entity recognition method based on Decision Tree. In 2002, Takeuchi and Collier [15] used Support Vector Machine-based approach to identify entities in the field of molecular biology. In 2003, De Meulder and Daelemans [16] proposed a named entity recognition method based on Memory Learning Algorithm.

Although the statistical-based methods are effective, they need well-selected features and have to go through more complex training. Moreover, these methods rely heavily on the corpus, but the labelled corpus that can be used to construct and evaluate the named entity recognition system is relatively small [3].

4 Hybrid Method

The hybrid methods [3] generally perform filtering and pruning through a set of rules in advance based on traditional models.

In 2006, Zhou et al. [17] proposed Chinese named entity recognition based on Cascaded Conditional Random Field Model (CCRF). The authors first used the N-shortest path segmentation strategy to perform text segmentation. Then they used the underlying CRF model to identify ordinary and non-nested names and placed names in the segmented result set. Finally, a high-level CRF model was used to identify the institution name. The recall rate of the model was 90.05%, and the accuracy rate was 88.12%. The performance was better than other Chinese institution name recognition algorithms. However, the over-fitting problem can arise.

In addition, in 2006, Yu et al. [18] proposed a Chinese named entity recognition based on Cascading Hidden Markov Model (CHMM) by using the method of internal cascading fusion of statistical learning methods. In 2009, Liao and Veeramachaneni [19] proposed a semi-supervised learning algorithm based on a conditional random field model for NER. The semi-supervised learning algorithm utilized unlabelled data to mitigate the impact of insufficient labelled data.

Since natural language processing is not entirely a random process, the use of statistical-based methods alone makes the state search space extremely large, and it is necessary to perform filtering and pruning in advance with the help of rules [3]. At present, almost all NER systems use statistical models based on rule knowledge.

5 Deep Learning Based Approach

In recent years, with the rapid expansion of the scale of Internet text data and the rapid development of deep learning technology, a large number of new methods for NER based on deep neural networks have emerged, which greatly improves the recognition accuracy. The models include: Recurrent Neural Network Model (RNN), Convolutional Neural Networks Model (CNN), Long Short-Term Memory Model (LSTM), LSTM-CRF Model, etc.

In 2011, Collobert et al. [20] used neural networks for NER, and finally added the CRF model to the objective function (later generally referred to as: a layer of CRF layer was combined).

Drawing on the above CRF ideas, a series of work combining the RNN structure and the CRF layer for NER appeared around 2015. Huang et al. [21] proposed a series of sequence labelling models based on LSTM model, including LSTM model, bidirectional LSTM model (BI-LSTM), LSTM model with CRF layer (LSTM-CRF) and BI-LSTM model with CRF layer (BI-LSTM-CRF). The BI-LSTM-CRF model first feeded the words into the BI-LSTM model, and then feeded all the scores predicted by the BI-LSTM block to the CRF layer, and finally selected the tag sequence with the highest predicted score as the best labels. Its accuracy was as high as 94.27% and 97.46%, respectively. The experimental results showed that the BI-LSTM-CRF model has reached or exceeded the CRF model and inherited the advantages of the deep learning method without the need for a large number of manually defined features.

Compared with traditional models, the deep learning based methods can embrace the context information and can avoid the laborious work of feature selection in models like CRF. The model obtained by further stacking the CRF layer over the neural network can achieve better results. However, methods based on deep neural network generally require larger training data.

6 Latest Method

Recently, in the research of neural-network-based named entity recognition, there are two new prevalent trends worth of noticing: one is the use of Attention Mechanism; the other is models based on a small number of labeled training data, including Transfer Learning and Semi-Supervised Learning.

6.1 Attention Mechanism

The Attention mechanism was first proposed in the field of visual images, and its essence comes from the human visual attention mechanism.

In 2014, Bahdanau et al. [22] used a similar attention-based mechanism to simultaneously translated and aligned on machine translation tasks. This work was currently recognized as the first to propose the application of the attention mechanism to the NLP field. The attention in deep learning can be interpreted broadly as a vector of importance weights.

In 2016, Rei et al. [23] focused on the splicing of word vectors and character vectors based on the RNN-CRF model structure. The algorithm first decomposed the word into single characters and got the corresponding character embeddings $(c_1, \ldots, c_R)$ which were feed into the BI-LSTM model, and then used the resulting hidden vector as the input word to obtain the corresponding vector m; and finally obtained the weighted sum by adding the vector m and the word embedding x together. Its weight was predicted by the hidden layer of traditional neural networks. The experimental results showed that the model is better than the original splicing method, the F1 values in the CoNLL-2003 dataset were 84.09% and 83.37%, respectively.

Introducing new parameters to compensate for the fitting ability of a certain aspect may not be comparable to the original method, and it may cause over-fitting and increase computational complexity.

6.2 Transfer Learning

The initial popular transfer learning in NLP was brought about by the term embedded model. An important challenge for sequence tagging is how to transfer knowledge from one task to another, which is often referred to as transfer learning [24]. Transfer learning can be used in several settings, notably for low-resource languages [25, 26] and low-resource domains such as biomedical corpora [27] and Twitter corpora [28].

In order to improve the performance on a target task by joint training with a source task, in 2017, Yang et al. [29] extended the basic sequence labelling model—NN-CRF model and proposed three migration learning architectures—T-A model, T-B model,

and T-C model. If the two domains had mutually mapped label sets, the algorithm proposed T-A model, that was, shared all model parameters and feature representations in the NN, and then performed a label mapping step at the top of the CRF layer to complete the cross-domain transfer. If the two domains had different label sets, then proposed a T-B model, which was to remove the parameter sharing in the CRF layer based on T-A model. At the same time, T-B model could also be adopted in the cross-domain migration learning. For cross-language migration, the algorithm proposed T-C model, which mainly performed migration learning by sharing word vectors and word-level layers between different languages.

The experimental results showed that the migration learning model proposed in the paper had significantly improved various data sets under low resource conditions and achieved better results in some benchmark tests.

6.3 Semi-supervised Learning

Marking the collected data wastes a lot of time and labour, and the unlabelled data can provide information about the distribution of the data, so we used this information to propose a semi-supervised algorithm. In the NLP field, the basic idea of semi-supervised learning is to use the model assumptions on the data distribution to establish a learner to label unlabelled samples.

In 2017, Peters et al. [30] proposed a semi-supervised method for sequence tagging tasks. The paper used a massive unlabelled corpus to train a bidirectional neural network language model (LM Embedding), and then used this model to obtain the language model vector of the current word to be annotated, and finally added the vector as a feature to the original bidirectional RNN-CRF model. The experimental results showed that the addition of this language model vector could greatly improve the sequence labelling performance on a small amount of annotation data.

7 Summary and Outlook

Named entity recognition has broad application prospects in many fields, such as biomedicine, medical text, finance, and so on. In the field of biomedicine, biomedical named entity recognition mainly identifies the named entities such as genes, proteins, disease names, drug names, and organization names in the biomedical literature [31]. In the field of medical texts, medical text entity recognition can fully exploit the value in information, and is an important basic work in medical knowledge mining, medical intelligent robots, medical clinical decision support systems and other application fields [32]. With the increasing international exchanges and the rapid development of the Internet, communication between different languages is becoming more and more important. Therefore, the application of named entity recognition in the field of machine translation also has a broader development space.

This paper introduces various existing methods for identifying named entities, including rules and dictionary based methods, HMM-based methods, CRF-based methods, deep learning-based methods, and so on. At present, the research on NER is relatively mature, but NER still has problems such as easy over-fitting, good effect in

individual fields, and inability to achieve versatility. At the same time, due to the large number of Chinese named entities, the complex structure and the ambiguity problem, the Chinese named entity recognition task more complicated than the English named entity recognition.

In recent years, the introduction of semi-supervised learning and unsupervised learning algorithms are trying to solve the problem of insufficient corpus. NER will integrate all aspects of development in a more open field and lay a solid foundation for the deep development of natural language processing.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (No. 61701284, No. 61702306, No. 61602278), Ministry of Education Humanities and Social Sciences Research Youth Fund Project (17YJCZH187) and Qingdao Philosophy, Social Science Planning Project (QDSKL1801131).

References

1. Chinchor N (1995) MUC-6 named entity task definition (version 2.1). In: Proceedings of the 6th conference on message understanding, Columbia, Maryland; Bakushinsky A, Goncharsky A (1994) Ill-posed problems theory and applications
2. Chinchor N, Robinson P (1997) MUC-7 named entity task definition. In: Proceedings of the 7th conference on message understanding, Columbia, Maryland
3. Sun Z, Wang H (2010) Overview on the advance of the research on name entity recognition. Data Anal Knowl Disc 26(6):42–47 (中文)
4. Rau LF (1991) Extracting company names from text. In: Proceedings of the seventh IEEE conference on artificial intelligence applications. IEEE
5. Zhang X, Wang L (1997) Identification and analysis of chinese organization and institution names. J Chin Inf Process 11(4):22–33 (中文)
6. Farmakiotou D, Karkaletsis V, Koutsias J et al (2000) Rule-based named entity recognition for Greek financial texts. In: Proceedings of the workshop on computational lexicography and multimedia dictionaries (COMLEX 2000), pp 75–78
7. Wang N, Ge R, Yuan C et al (2002) Company name identification in Chinese financial domain. Chin J Inf Sci 16(2):1–6 (中文)
8. Li H (2012) Statistical learning method. Tsinghua University Press (中文)
9. Bikel DM, Miller S, Schwartz R et al (1998) Nymble: a high-performance learning name-finder. arXiv preprint cmp-lg/9803003
10. Bikel DM, Schwartz R, Weischedel RM (1999) An algorithm that learns what's in a name. Mach Learn 34(1–3):211–231
11. Liu J (2009) Chinese named entity recognition algorithm based on improved hidden Markov model. J Taiyuan Normal Univ Nat Sci Ed 1:80–83 (中文)
12. Krishnan V, Manning CD (2006) Association for computational linguistics the 21st international conference, Sydney, Australia (2006.07.17–2006.07.18). Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the ACL, ACL'06—An effective two-stage model for exploiting non-local dependencies in named entity recognition. International conference on computational linguistics & the meeting of the association for computational linguistics. Association for Computational Linguistics, pp 1121–1128

13. Borthwick A, Sterling J, Agichtein E et al (1998) NYU: description of the MENE named entity system as used in MUC-7. In: Seventh message understanding conference (MUC-7): proceedings of a conference held in Fairfax, Virginia, April 29–May 1998
14. Sekine S, Grishman R, Shinnou H (1998) A decision tree method for finding and classifying names in Japanese texts. In: Sixth workshop on very large corpora
15. Takeuchi K, Collier N (2002) Use of support vector machines in extended named entity recognition, In: Proceedings of the 6th conference on natural language learning, vol 20. Association for Computational Linguistics, pp 1–7
16. De Meulder F, Daelemans W (2003) Memory-based named entity recognition using unannotated data. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, vol 4. Association for Computational Linguistics, pp 208–211
17. Zhou J, Dai X, Yin C et al (2006) Automatic recognition of Chinese organization name based on cascaded conditional random fields. Chin J Electron 34(5):804–809 (中文)
18. Yu H, Zhang H, Liu Q et al (2006) Automatic recognition of Chinese organization name based on cascaded conditional random fields. Trans Commun 2 (中文)
19. Liao W, Veeramachaneni S (2009) A simple semi-supervised algorithm for named entity recognition. In: Proceedings of the NAACL HLT 2009 workshop on semi-supervised learning for natural language processing. Association for Computational Linguistics, pp 58–65
20. Collobert R, Weston J, Bottou L et al (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12(Aug):2493–2537
21. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991
22. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
23. Rei M, Crichton GKO, Pyysalo S (2016) Attending to characters in neural sequence labeling models. arXiv preprint arXiv:1611.04361
24. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22 (10):1345–1359
25. Zirikly A, Hagiwara M (2015) Cross-lingual transfer of named entity recognizers without parallel corpora. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, vol 2: Short papers, pp 390–396
26. Wang M, Manning CD (2014) Cross-lingual projected expectation regularization for weakly supervised learning. Trans Assoc Comput Linguist 2:55–66
27. Kim JD, Ohta T, Tateisi Y et al (2003) GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl_1):i180–i182
28. Ritter A, Clark S, Etzioni O (2011) Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 1524–1534
29. Yang Z, Salakhutdinov R, Cohen WW (2017) Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345
30. Peters ME, Ammar W, Bhagavatula C et al (2017) Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108
31. Zheng Q, Liu Q, Wang Z et al (2010) Research and development on biomedical named entity recognition. J Comput Appl 27(3) (中文)
32. Zhang F, Wang M (2017) Medical text entities recognition method base on deep learning. Comput Technol Autom 36(1):123 (中文)